Journal of Mathematical Analysis and Applications **253,** 58–78 (2001)
doi:10.1006/jmaa.2000.7055, available online at http://www.idealibrary.com on IDEAL®

On Markov Policies for Minimax Decision Processes

Seiichi Iwamoto and Kazuyoshi Tsurusaki

Department of Economic Engineering, Faculty of Economics, Kyushu University 27

and

Toshiharu Fujita

Department of Mathematics, Graduate School of Mathematics, Kyushu University 33, Fukuoka 812-8581, Japan

Submitted by William F. Ames

Received September 24, 1996

1. INTRODUCTION

Since Bellman [1] originated, developed, and applied "dynamic programming," an enormous amount of efforts has been devoted to the study of both deterministic and stochastic dynamic programming ([4, 5, 8, 10–12, 22, 25–27, 29–31] and others). (As for papers, see also [6, 7, 13, 14, 18, 19, 21, 23, 24] and others.) As usual, deterministic dynamic programming is called dynamic programming. The stochastic dynamic programming is called the Markov decision process. This study is mainly concerned with the sequential optimization of *additive* function as objective function, which includes the discounted case [28]. In this paper we call this problem the *additive problem*. It is tacitly known that there exists an optimal policy which is Markov for both additive problems ([3, p. 152; 5, p. 6], and others). In fact, some papers are restricted at the outset to the set of all Markov policies. And then they have tried to find an optimal policy for the problems under consideration. However, first of all, it should be clarified that the plausibility of this restriction is reasonable. Sometimes, for some reason or other, the clarification is omitted.

In this paper, we are concerned with stochastic maximization problems of *minimum* function, called the *minimum problem*. We raise the question

0022-247X/01 $35.00

of whether there exists an optimal policy for the stochastic minimum or not. Furthermore, if it exists, we focus our attention on the question of whether the optimal policy is Markov or not.

Section 2 discusses stochastic maximization of the minimum function. Section 3 discusses an imbedded problem of the stochastic maximization problem. An optimal (not necessarily Markov) policy is derived through the invariant imbedding approach [2, 15, 20]. The last section illustrates a two-stage stochastic decision process which does not admit any optimal Markov policy. This is verified both by a brute force enumeration method and by the invariant imbedding method.

Throughout the paper we use the following notations:

$$
\begin{aligned}
&N \geq 2 \text{ is an integer; the total number of stages} \\
&X = \{s_1, s_2, \ldots, s_p\} \text{ is a finite state space} \\
&U = \{a_1, a_2, \ldots, a_k\} \text{ is a finite action space} \\
&r_n \colon X \times U \to R^1 \text{ is an } n\text{th reward function} \qquad (0 \leq n \leq N-1) \\
&r_G \colon X \to R^1 \text{ is a terminal reward function} \\
&p \text{ is a Markov transition law:} \\
&\qquad p(y \mid x, u) \geq 0 \qquad \forall (x, u, y) \in X \times U \times X, \\
&\qquad \sum_{y \in X} p(y \mid x, u) = 1 \qquad \forall (x, u) \in X \times U
\end{aligned}
\tag{1}
$$

$y \sim p(\cdot \mid x, u)$ denotes that next state y conditioned on state x and action u appears with probability $p(y \mid x, u)$.

2. STOCHASTIC MAXIMIZATION OF MINIMUM FUNCTION

Let us consider the stochastic maximization problem with minimum function,

$$
\begin{aligned}
\text{Maximize} \quad & E\big[r_0(x_0, u_0) \wedge r_1(x_1, u_1) \wedge \cdots \\
& \qquad \wedge r_{N-1}(x_{N-1}, u_{N-1}) \wedge r_G(x_N)\big] \\
\text{subject to} \quad & \text{(i)} \quad x_{n+1} \sim p(\cdot \mid x_n, u_n) \\
& \text{(ii)} \quad u_n \in U \qquad n = 0, 1, \ldots, N-1.
\end{aligned}
\tag{2}
$$

2.1. *General Policies*

In this subsection we consider the problem (2) with the set of all general policies, called the *general problem*. Any general policy $\sigma = \{\sigma_n, \ldots, \sigma_{N-1}\}$

over the $(N-n)$-stage process yields its expected value,

$$J^n(x_n;\sigma) = \sum_{(x_{n+1},\ldots,x_N)\in X\times\cdots\times X}\sum\cdots\sum \{[r_n(x_n,u_n) \wedge \cdots \wedge r_{N-1}(x_{N-1},u_{N-1}) \wedge r_G(x_N)] \times p(x_{n+1}|x_n,u_n)\cdots p(x_N|x_{N-1},u_{N-1})\}, \quad (3)$$

where $\{u_n, x_{n+1},\ldots,x_{N-1},u_{N-1},x_N\}$ is stochastically generated through the general policy σ and the starting state x_n as

$$\begin{aligned}\sigma_n(x_n) = u_n &\to p(\cdot|x_n,u_n) \sim x_{n+1}\\ &\to \sigma_{n+1}(x_n,x_{n+1}) = u_{n+1} \to p(\cdot|x_{n+1},u_{n+1}) \sim x_{n+2}\\ &\to \sigma_{n+2}(x_n,x_{n+1},x_{n+2}) = u_{n+2} \to p(\cdot|x_{n+2},u_{n+2}) \sim x_{n+3} \to \cdots\\ &\to \sigma_{N-1}(x_n,x_{n+1},\ldots,x_{N-1}) = u_{N-1} \to p(\cdot|x_{N-1},u_{N-1}) \sim x_N.\end{aligned}$$

We define the following family of *general subproblems*:

$$\begin{aligned} V^N(x_N) &= r_G(x_N), && x_N \in X\\ V^n(x_n) &= \max_{\sigma} J^n(x_n;\sigma), && x_n \in X, \quad 0 \le n \le N-1. \end{aligned} \quad (4)$$

Thus the general problem (2) is identical to (4) with $n=0$. However, in general, the recursive formula for the general subproblems,

$$\begin{aligned} V^N(x) &= r_G(x), && x \in X\\ V^n(x) &= \max_{u\in U}\left[r_n(x,u) \wedge \sum_{y\in X} V^{n+1}(y)\,p(y|x,u)\right], && x\in X, \quad 0\le n\le N-1, \end{aligned}$$

does not hold.

Nevertheless, we have the following positive result:

THEOREM 2.1. *A general policy yields the optimal value function $V^0(\cdot)$ for the general problem. That is, there exists an optimal general policy σ^* for the general problem* (2):

$$J^0(x_0;\sigma^*) = V^0(x_0) \qquad \text{for all } x_0 \in X.$$

In fact, an invariant imbedding approach [2, 15, 20] for the general problem (2) yields an optimal general policy $\sigma^* = \{\sigma_0^*,\ldots,\sigma_{N-1}^*\}$. This we proceed to construct in Section 3.3.

2.2. *Markov Policies*

In this subsection we consider the problem (2) restricted to the set of all Markov policies, as Bellman and Zadeh [3, Sect. 5] have done. We call this problem the *Markov problem*. Any Markov policy $\pi = \{\pi_n, \ldots, \pi_{N-1}\}$ over the $(N-n)$-stage process yields its expected value $J^n(x_n; \pi)$, where

$$\pi_n(x_n) = u_n,\ \pi_{n+1}(x_{n+1}) = u_{n+1}\ , \ldots,\ \pi_{N-1}(x_{N-1}) = u_{N-1}.$$

We define the corresponding *Markov subproblems* as

$$\begin{aligned} v^N(x_N) &= r_G(x_N), && x_N \in X \\ v^n(x_n) &= \operatorname*{Max}_{\pi} J^n(x_n; \pi), && x_n \in X,\quad 0 \le n \le N-1. \end{aligned} \tag{5}$$

Then the Markov problem (2) becomes (5) with $n = 0$. In general, the recursive formula for the Markov subproblems,

$$\begin{aligned} v^N(x) &= r_G(x), && x \in X \\ v^n(x) &= \operatorname*{Max}_{u \in U}\left[r_n(x,u) \wedge \sum_{y \in X} v^{n+1}(y)\, p(y \mid x, u)\right], \\ & && x \in X,\quad 0 \le n \le N-1, \end{aligned} \tag{6}$$

does not hold. We remark that Bellman and Zadeh derive the recursive formula for $\{v^0(\cdot), v^1(\cdot), \ldots, v^N(\cdot)\}$ [3, Sect. 5]. (See also [9, 17].) However, the recursive formula (6) does not hold, as shown by Iwamoto and Fujita [16].

THEOREM 2.2. *In general, Markov policy does not yield the optimal value function $V^0(\cdot)$ for the general problem. That is, there exists a stochastic decision process with minimum function such that for any Markov policy π,*

$$V^0(x_0) > J^0(x_0; \pi) \qquad \textit{for some } x_0 \in X.$$

Proof. The proof will be completed by illustrating an example in Section 4.1. ∎

3. IMBEDDED PROCESS WITH MINIMUM FUNCTION

Let us return to the original stochastic maximization problem (2) with minimum function. Note that, without loss of generality, we may assume that

$$\begin{aligned} &0 \le r_n(x,u) \le 1 && (x,u) \in X \times U,\quad 0 \le n \le N-1 \\ &0 \le r_G(x) \le 1 && x \in X. \end{aligned} \tag{7}$$

In this section, under the condition (7), we imbed the problem (2) into the family of parameterized problems,

$$\begin{aligned} &\text{Maximize} \quad E\big[\lambda_0 \wedge r_0(x_0, u_0) \wedge r_1(x_1, u_1) \wedge \cdots \\ &\qquad\qquad \wedge r_{N-1}(x_{N-1}, u_{N-1}) \wedge r_G(x_N)\big] \\ &\text{subject to} \quad \text{(i)} \quad x_{n+1} \sim p(\cdot \mid x_n, u_n) \\ &\qquad\qquad \text{(ii)} \quad u_n \in U, \qquad n = 0, 1, \ldots, N-1, \end{aligned} \tag{8}$$

where the parameter ranges over $\lambda_0 \in [0, 1]$.

3.1. *General Policies*

First we consider the imbedded problem (8) with the set of all general policies, called the *general problem*. Here we note that any general policy,

$$\sigma = \{\sigma_0, \sigma_1, \ldots, \sigma_{N-1}\},$$

consists of the decision functions

$$\begin{aligned} \sigma_0&: \quad X \times [0,1] \to U \\ \sigma_1&: \quad (X \times [0,1]) \times (X \times [0,1]) \to U \\ &\cdots \\ \sigma_{N-1}&: \quad (X \times [0,1]) \times (X \times [0,1]) \times \cdots \times (X \times [0,1]) \to U. \end{aligned}$$

Thus, any general policy $\sigma = \{\sigma_n, \ldots, \sigma_{N-1}\}$ over the $(N-n)$-stage process yields its expected value,

$$\begin{aligned} K^n(x_n, \lambda_n; \sigma) = \sum\sum\cdots\sum_{(x_{n+1}, \ldots, x_N) \in X \times \cdots \times X} &\big\{\big[\lambda_n \wedge r_n(x_n, u_n) \wedge \cdots \\ &\wedge r_{N-1}(x_{N-1}, u_{N-1}) \wedge r_G(x_N)\big] \\ &\times p(x_{n+1} \mid x_n, u_n) \cdots p(x_N \mid x_{N-1}, u_{N-1})\big\}, \end{aligned} \tag{9}$$

where the alternating sequence of action and two-dimensional state,

$$\{u_n, (x_{n+1}, \lambda_{n+1}), u_{n+1}, (x_{n+2}, \lambda_{n+2}), \ldots, u_{N-1}, (x_N, \lambda_N)\},$$

is stochastically generated through the policy σ and the starting state

(x_n, λ_n) as

$$\sigma_n(x_n, \lambda_n) = u_n \to \begin{cases} p(\cdot | x_n, u_n) \sim x_{n+1} \\ \lambda_n \wedge r_n(x_n, u_n) = \lambda_{n+1} \end{cases}$$

$$\to \sigma_{n+1}(x_n, \lambda_n, x_{n+1}, \lambda_{n+1}) = u_{n+1}$$

$$\to \begin{cases} p(\cdot | x_{n+1}, u_{n+1}) \sim x_{n+2} \\ \lambda_{n+1} \wedge r_{n+1}(x_{n+1}, u_{n+1}) = \lambda_{n+2} \end{cases}$$

$$\to \sigma_{n+2}(x_n, \lambda_n, x_{n+1}, \lambda_{n+1}, x_{n+2}, \lambda_{n+2}) = u_{n+2}$$

$$\to \begin{cases} p(\cdot | x_{n+2}, u_{n+2}) \sim x_{n+3} \\ \lambda_{n+2} \wedge r_{n+2}(x_{n+2}, u_{n+2}) = \lambda_{n+3} \end{cases} \to \cdots$$

$$\to \sigma_{N-1}(x_n, \lambda_n, x_{n+1}, \lambda_{n+1}, \ldots, x_{N-1}, \lambda_{N-1}) = u_{N-1}$$

$$\to \begin{cases} p(\cdot | x_{N-1}, u_{N-1}) \sim x_N \\ \lambda_{N-1} \wedge r_{N-1}(x_{N-1}, u_{N-1}) = \lambda_N. \end{cases} \tag{10}$$

However, note that the sequence of the latter halves of the states $\{\lambda_{n+1}, \lambda_{n+2}, \ldots, \lambda_N\}$ behaves deterministically.

We define the family of the corresponding *general subproblems*:

$$V^N(x_N, \lambda_N) = \lambda_N \wedge r_G(x_N), \qquad x_N \in X, \quad 0 \le \lambda_N \le 1$$

$$V^n(x_n, \lambda_n) = \operatorname*{Max}_{\sigma} K^n(x_n, \lambda_n; \sigma), \tag{11}$$

$$x_N \in X, \quad 0 \le \lambda_n \le 1, \quad 0 \le n \le N-1.$$

Then the general problem (8) is identical to (11) with $n = 0$. We have the recursive formula for the general subproblems:

THEOREM 3.1.

$$V^N(x, \lambda) = \lambda \wedge r_G(x), \qquad x \in X, \lambda \in [0,1]$$

$$V^n(x, \lambda) = \operatorname*{Max}_{u \in U} \sum_{y \in X} V^{n+1}(y, \lambda \wedge r_n(x,u)) p(y | x, u), \tag{12}$$

$$x \in X, \lambda \in [0,1], \quad 0 \le n \le N-1.$$

3.2. *Markov Policies*

Next we consider the *Markov problem*. That is, we restrict the imbedded problem (8) to the set of all Markov policies. Here Markov policy

$$\pi = \{\pi_0, \pi_1, \ldots, \pi_{N-1}\}$$

consists in turn of two-variable decision functions:

$$\pi_n\colon X \times [0,1] \to U, \quad 0 \le n \le N-1.$$

Note that any Markov policy $\pi = \{\pi_n, \ldots, \pi_{N-1}\}$ over the $(N-n)$-stage process yields its expected value $K^n(x_n, \lambda_n; \pi)$ through (9). The alternating sequence of action and two-dimensional state

$$\{u_n, (x_{n+1}, \lambda_{n+1}), u_{n+1}, (x_{n+2}, \lambda_{n+2}), \ldots, u_{N-1}, (x_N, \lambda_N)\}$$

is similarly generated through the policy π and the state (x_n, λ_n) as in (10), where

$$\begin{aligned}
&\pi_n(x_n, \lambda_n) = u_n \\
&\pi_{n+1}(x_{n+1}, \lambda_{n+1}) = u_{n+1} \\
&\cdots \\
&\pi_{N-1}(x_{N-1}, \lambda_{N-1}) = u_{N-1}.
\end{aligned}$$

Of course, the sequence of the latter halves of the states $\{\lambda_{n+1}, \lambda_{n+2}, \ldots, \lambda_N\}$ behaves deterministically.

We define the family of the corresponding *Markov subproblems*:

$$\begin{aligned}
v^N(x_N, \lambda_N) &= \lambda_N \wedge r_G(x_N), \quad x_N \in X, \quad 0 \le \lambda_N \le 1 \\
v^n(x_n, \lambda_n) &= \operatorname*{Max}_{\pi} K^n(x_n, \lambda_n; \pi) \qquad (13) \\
&\qquad x_n \in X, \quad 0 \le \lambda_n \le 1, \quad 0 \le n \le N-1.
\end{aligned}$$

Note that the Markov problem (8) is also (13) with $n = 0$. Then we have the recursive formula for the Markov subproblems:

THEOREM 3.2.

$$\begin{aligned}
v^N(x, \lambda) &= \lambda \wedge r_G(x), \qquad x \in X, \ \lambda \in [0,1] \\
v^n(x, \lambda) &= \operatorname*{Max}_{u \in U} \sum_{y \in X} v^{n+1}(y, \lambda \wedge r_n(x,u)) p(y \mid x, u) \qquad (14) \\
&\qquad x \in X, \ \lambda \in [0,1], \quad 0 \le n \le N-1.
\end{aligned}$$

THEOREM 3.3. (i) *A Markov policy yields the optimal value function* $V^0(\cdot)$ *for the general problem. That is, there exists an optimal Markov policy* π^* *for the general problem* (8):

$$V^0(x_0, \lambda_0) = K^0(x_0, \lambda_0; \pi^*) \qquad \textit{for all } (x_0, \lambda_0) \in X \times [0,1].$$

In fact, letting $\pi_n^*(x, \lambda)$ *be a maximizer of* (14) (*or* (12)) *for each* $(x, \lambda) \in X \times [0,1]$, $0 \le n \le N-1$, *we have the optimal Markov policy* $\pi^* = \{\pi_0^*, \ldots, \pi_{N-1}^*\}$.

(ii) *The optimal value functions for the Markov subproblems* (13) *are equal to the optimal value functions for the general problems* (11):

$$v^n(x, \lambda) = V^n(x, \lambda), \qquad (x, \lambda) \in X \times [0,1], \quad 0 \le n \le N.$$

3.3. *Proof of Theorem* 2.1

Now, in this subsection, let us prove Theorem 2.1 by use of the result of Theorem 3.3.

First we note that any Markov policy for the imbedded problem (8) $\pi = \{\pi_0, \ldots, \pi_{N-1}\}$ together with a specified value of the parameter λ_0 induces the general policy for the problem (2) $\sigma = \{\sigma_0, \ldots, \sigma_{N-1}\}$ as

$$\sigma_0(x_0) := \pi_0(x_0, \lambda_0)$$
$$\sigma_1(x_0, x_1) := \pi_1(x_1, \lambda_1)$$

where

$$\lambda_1 = \lambda_0 \wedge r_0(x_0, u_0), \qquad u_0 = \pi_0(x_0, \lambda_0)$$
$$\sigma_2(x_0, x_1, x_2) := \pi_2(x_2, \lambda_2)$$

where

$$\begin{aligned} &\lambda_2 = \lambda_1 \wedge r_1(x_1, u_1), \qquad u_1 = \pi_1(x_1, \lambda_1), \qquad \lambda_1 = \lambda_0 \wedge r_0(x_0, u_0), \\ &u_0 = \pi_0(x_0, \lambda_0) \\ &\cdots \\ &\sigma_{N-1}(x_0, x_1, \ldots, x_{N-1}) := \pi_{N-1}(x_{N-1}, \lambda_{N-1}) \end{aligned} \tag{15}$$

where

$$\begin{aligned} &\lambda_{N-1} = \lambda_{N-2} \wedge r_{N-2}(x_{N-2}, u_{N-2}), \qquad u_{N-2} = \pi_{N-2}(x_{N-2}, \lambda_{N-2}), \\ &\lambda_{N-2} = \lambda_{N-3} \wedge r_{N-3}(x_{N-3}, u_{N-3}), \qquad u_{N-3} = \pi_{N-3}(x_{N-3}, \lambda_{N-3}), \ldots, \\ &\lambda_1 = \lambda_0 \wedge r_0(x_0, u_0), \qquad u_0 = \pi_0(x_0, \lambda_0). \end{aligned}$$

Furthermore, we see that the Markov policy π with a specified value $\lambda_0 = 1$ and the resulting general policy σ yield the same value function:

$$K^0(x_0, 1; \pi) = J^0(x_0; \sigma), \qquad x_0 \in X.$$

Second we note that Theorem 3.3 ensures the existence of an optimal Markov policy for the imbedded problem (8) π; which together with the

value $\lambda_0 = 1$ induces the corresponding general policy for the problem (2) σ^*, as is shown by (15). Thus we get

$$K^0(x_0, 1; \pi^*) = J^0(x_0; \sigma^*), \qquad x_0 \in X. \tag{16}$$

On the other hand, for any general policy for the problem (2) $\sigma = \{\sigma_0, \ldots, \sigma_{N-1}\}$ we define a general policy for the imbedded problem (8) $\tilde{\sigma} = \{\tilde{\sigma}_0, \ldots, \tilde{\sigma}_{N-1}\}$ by

$$\tilde{\sigma}_n(x_0, \lambda_0, x_1, \lambda_1, \ldots, x_n, \lambda_n) := \sigma_n(x_0, x_1, \ldots, x_n)$$

on

$$(X \times [0,1]) \times (X \times [0,1]) \times \cdots \times (X \times [0,1]), \quad 0 \le n \le N-1.$$

Then we have

$$K^0(x_0, 1; \tilde{\sigma}) = J^0(x_0; \sigma), \qquad x_0 \in X. \tag{17}$$

Therefore, the optimality of the policy π^* implies

$$K^0(x_0, 1; \pi^*) \ge K^0(x_0, 1; \tilde{\sigma}), \qquad x_0 \in X. \tag{18}$$

Combining (16), (17), and (18), we get for any general policy σ

$$\begin{aligned} J^0(x_0; \sigma^*) &= K^0(x_0, 1; \pi^*) \\ &\ge K^0(x_0, 1; \tilde{\sigma}) \\ &= J^0(x_0; \sigma). \end{aligned}$$

Thus the policy σ^* is optimal for the general problem (2). This completes the proof of Theorem 2.1.

3.4. *Proofs of Theorems* 3.1–3.3

In this subsection we prove only Theorems 3.1 and 3.3(i), because Theorems 3.2 and 3.3(ii) are the direct consequences of Theorems 3.1 and 3.3(i). We prove both theorems for the two-stage process, because the theorems for the N-stage process are proved similarly.

We note that for $(x_n, \lambda_n) \in X \times [0,1]$,

$$V^2(x_2, \lambda_2) = \lambda_2 \wedge r_G(x_2)$$

$$V^1(x_1, \lambda_1) = \operatorname*{Max}_{\sigma_1} \sum_{x_2 \in X} \left[\lambda_1 \wedge r_1(x_1, u_1) \wedge r_G(x_2)\right] p(x_2 \mid x_1, u_1) \tag{19}$$

$$V^0(x_0, \lambda_0) = \operatorname*{Max}_{\sigma_0, \sigma_1} \sum\sum_{(x_1, x_2) \in X \times X} \left\{\left[\lambda_0 \wedge r_0(x_0, u_0) \wedge r_1(x_1, u_1) \wedge r_G(x_2)\right] \times p(x_1 \mid x_0, u_0)\, p(x_2 \mid x_1, u_1)\right\}, \tag{20}$$

where $u_1 = \sigma_1(x_1, \lambda_1)$ in (19) and $u_0 = \sigma_0(x_0, \lambda_0)$, $\lambda_1 = \lambda_0 \wedge r_0(x_0, u_0)$, $u_1 = \sigma_1(x_0, \lambda_0, x_1, \lambda_1)$ in (20), respectively.

Thus the equality

$$V^1(x_1, \lambda_1) = \operatorname*{Max}_{u_1 \in U} \sum_{x_2 \in X} V^2\big(x_2, \lambda_1 \wedge r_1(x_1, u_1)\big) p(x_2 \mid x_1, u_1),$$

$$x_1 \in X,\ \lambda_1 \in [0,1]$$

is trivial. We prove

$$V^0(x_0, \lambda_0) = \operatorname*{Max}_{u_0 \in U} \sum_{x_1 \in X} V^1\big(x_1, \lambda_0 \wedge r_0(x_0, u_0)\big) p(x_1 \mid x_0, u_0),$$

$$x_0 \in X,\ \lambda_0 \in [0,1]. \quad (21)$$

Let us choose an optimal (Markov) policy σ_1^* for the one-stage process:

$$V^1(x_1, \lambda_1) = K^1(x_1, \lambda_1; \sigma_1^*) \qquad \forall (x_1, \lambda_1) \in X \times [0,1]. \quad (22)$$

From the definition (11), we can choose for each $(x_0, \lambda_0) \in X \times [0,1]$ an optimal (not necessarily Markov) policy $\tilde{\sigma} = \{\tilde{\sigma}_0, \tilde{\sigma}_1\}$ for the two-stage process:

$$V^0(x_0, \lambda_0) = K^0(x_0, \lambda_0; \tilde{\sigma}) \qquad (x_0, \lambda_0) \in X \times [0,1].$$

Thus we see that

$$V^0(x_0, \lambda_0) = \sum\sum_{(x_1, x_2) \in X \times X} \big\{\big[\lambda_0 \wedge r_0(x_0, u_0) \wedge r_1(x_1, u_1) \wedge r_G(x_2)\big] \times p(x_1 \mid x_0, u_0) p(x_2 \mid x_1, u_1)\big\}, \quad (23)$$

where

$$u_0 = \tilde{\sigma}_0(x_0, \lambda_0), \qquad \lambda_1 = \lambda_0 \wedge r_0(x_0, u_0), \qquad u_1 = \tilde{\sigma}_1(x_0, \lambda_0, x_1, \lambda_1). \quad (24)$$

We note that

$$\sum\sum_{(x_1, x_2) \in X \times X} f(x_1, x_2) = \sum_{x_1 \in X} \sum_{x_2 \in X} f(x_1, x_2) \quad (25)$$

and

$$\sum_{x_2 \in X} \big[\lambda_1 \wedge r_1(x_1, u_1) \wedge r_G(x_2)\big] p(x_2 \mid x_1, u_1) \le K^1(x_1, \lambda_1; \sigma_1^*) = V^1(x_1, \lambda_1) \qquad \forall (x_1, \lambda_1) \in X \times [0,1]. \quad (26)$$

From (23), together with (24), (25), and (26), we have

$$V^0(x_0, \lambda_0) \le \sum_{x_1 \in X} \sum_{x_2 \in X} \{[\lambda_0 \wedge r_0(x_0, u_0) \wedge r_1(x_1, u_1) \wedge r_G(x_2))] \times p(x_1 | x_0, u_0) p(x_2 | x_1, u_1)\}$$

$$= \sum_{x_1 \in X} \{ \sum_{x_2 \in X} \{[\lambda_1 \wedge r_1(x_1, u_1) \wedge r_G(x_2)] p(x_2 | x_1, u_1)\} \times p(x_1 | x_0, u_0)\} \qquad (\lambda_1 = \lambda_0 \wedge r_0(x_0, u_0))$$

$$\le \sum_{x_1 \in X} V^1(x_1, \lambda_1) p(x_1 | x_0, u_0), \qquad (\lambda_1 = \lambda_0 \wedge r_0(x_0, u_0))$$

$$= \sum_{x_1 \in X} V^1(x_1, \lambda_0 \wedge r_0(x_0, u_0)) p(x_1 | x_0, u_0).$$

Consequently, we have

$$V^0(x_0, \lambda_0) \le \sum_{x_1 \in X} V^1(x_1, \lambda_0 \wedge r_0(x_0, u_0)) p(x_1 | x_0, u_0) \qquad \forall (x_0, \lambda_0) \in X \times [0, 1].$$

Thus taking the maximum over $u \in U$, we get

$$V^0(x_0, \lambda_0) \le \operatorname*{Max}_{u_0 \in U} \sum_{x_1 \in X} V^1(x_1, \lambda_0 \wedge r_0(x_0, u_0)) p(x_1 | x_0, u_0) \qquad \forall (x_0, \lambda_0) \in X \times [0, 1]. \quad (27)$$

On the other hand, for any $(x_0, \lambda_0) \in X \times [0, 1]$ let $u^* = u^*(x_0, \lambda_0) \in U$ be a maximizer of the right-hand side of (27). This defines a Markov decision function,

$$\pi_0^* : X \times [0, 1] \to U \qquad \pi_0^*(x_0, \lambda_0) = u^*(x_0, \lambda_0).$$

Then we have

$$\operatorname*{Max}_{u_0 \in U} \sum_{x_1 \in X} V^1(x_1, \lambda_0 \wedge r_0(x_0, u_0)) p(x_1 | x_0, u_0) = \sum_{x_1 \in X} V^1(x_1, \lambda_0 \wedge r_0(x_0, u_0)) p(x_1 | x_0, u_0) \qquad (u_0 = \pi_0^*(x_0, \lambda_0)). \quad (28)$$

From (22), we get

$$V^1(x_1, \lambda_1) = \sum_{x_2 \in X} \left[\lambda_1 \wedge r_1(x_1, u_1) \wedge r_G(x_2)\right] p(x_2 \mid x_1, u_1) \qquad (u_1 = \sigma_1^*(x_1, \lambda_1)). \quad (29)$$

Thus we have from (29)

$$\begin{aligned} &\sum_{x_1 \in X} V^1(x_1, \lambda_0 \wedge r_0(x_0, u_0)) p(x_1 \mid x_0, u_0) \qquad (u_0 = \pi_0^*(x_0, \lambda_0)) \\ &\quad = \sum_{x_1 \in X} \Big\{ \sum_{x_2 \in X} \left[\lambda_1 \wedge r_1(x_1, u_1) \wedge r_G(x_2)\right] p(x_2 \mid x_1, u_1) \Big\} p(x_1 \mid x_0, u_0) \\ &\qquad\qquad (\text{for } \lambda_1 = \lambda_0 \wedge r_0(x_0, u_0)) \\ &\quad = \sum\sum_{(x_1, x_2) \in X \times X} \Big\{ \left[\lambda_0 \wedge r_0(x_0, u_0) \wedge r_1(x_1, u_1) \wedge r_G(x_2)\right] \\ &\qquad\qquad \times p(x_1 \mid x_0, u_0) p(x_2 \mid x_1, u_1) \Big\}. \qquad (30) \end{aligned}$$

Combining (28) and (30), we obtain

$$\begin{aligned} &\operatorname*{Max}_{u_0 \in U} \sum_{x_1 \in X} V^1(x_1, \lambda_0 \wedge r_0(x_0, u_0)) p(x_1 \mid x_0, u_0) \\ &\quad = \sum\sum_{(x_1, x_2) \in X \times X} \Big\{ \left[\lambda_0 \wedge r_0(x_0, u_0) \wedge r_1(x_1, u_1) \wedge r_G(x_2)\right] \\ &\qquad\qquad \times p(x_1 \mid x_0, u_0) p(x_2 \mid x_1, u_1) \Big\} \\ &\qquad\qquad (u_0 = \pi_0^*(x_0, \lambda_0),\ \lambda_1 = \lambda_0 \wedge r_0(x_0, u_0),\ u_1 = \sigma_1^*(x_1, \lambda_1)) \\ &\quad \le \operatorname*{Max}_{\pi_0, \pi_1} \sum\sum_{(x_1, x_2) \in X \times X} \Big\{ \left[\lambda_0 \wedge r_0(x_0, u_0) \wedge r_1(x_1, u_1) \wedge r_G(x_2)\right] \\ &\qquad\qquad \times p(x_1 \mid x_0, u_0) p(x_2 \mid x_1, u_1) \Big\} \\ &\quad = V^0(x_0, \lambda_0). \qquad (31) \end{aligned}$$

Both equations (27) and (31) imply the desired equality (21). This completes the proof of Theorem 3.1.

Furthermore, the equalities in (31) imply that the optimal value function $V^0(\cdot)$ is attained by the Markov policy $\bar{\pi} = \{\pi_0^*, \sigma_1^*\}$:

$$V^0(x_0, \lambda_0) = K^0(x_0, \lambda_0; \bar{\pi}) \qquad (x_0, \lambda_0) \in X \times [0, 1].$$

This completes the proof of Theorem 3.3(i).

4. EXAMPLE

In this section we illustrate a stochastic decision process with minimum function which does not admit any optimal *Markov* policy. As was mentioned in Section 2.2, the illustration also proves Theorem 2.2.

We consider the two-stage, three-state, and two-action problem (2) as

$$\begin{array}{lll} \text{Maximize} & E\left[r_0(u_0) \wedge r_1(u_1) \wedge r_G(x_2)\right] & \\ \text{subject to} & \text{(i)} \quad x_{n+1} \sim p(\cdot|x_n, u_n) & n = 0, 1 \\ & \text{(ii)} \quad u_0 \in U, \quad u_1 \in U, & \end{array} \tag{32}$$

where the data are given as follows (see also [3, p. B154, 16]):

$$r_G(s_1) = 0.5 \qquad r_G(s_2) = 0.2 \qquad r_G(s_3) = 0.8$$
$$r_1(a_1) = 1.0 \qquad r_1(a_2) = 0.8$$
$$r_0(a_1) = 0.9 \qquad r_0(a_2) = 0.6$$

$u_t = a_1$

$x_t \backslash x_{t+1}$	s_1	s_2	s_3
s_1	0.4	0.5	0.1
s_2	0.2	0.6	0.2
s_3	0.3	0.1	0.6

$u_t = a_2$

$x_t \backslash x_{t+1}$	s_1	s_2	s_3
s_1	0.1	0.6	0.3
s_2	0.7	0.2	0.1
s_3	0.3	0.3	0.4

4.1. *Brute Force Enumeration Method*

In this subsection we solve the problem directly by generating two-stage stochastic decision trees and enumerating all of the possible histories together with the related expected values.

We remark that the size yields $2^3 = 8$ first decision functions,

$$\sigma_0 = \begin{pmatrix} \sigma_0(s_1) \\ \sigma_0(s_2) \\ \sigma_0(s_3) \end{pmatrix},$$

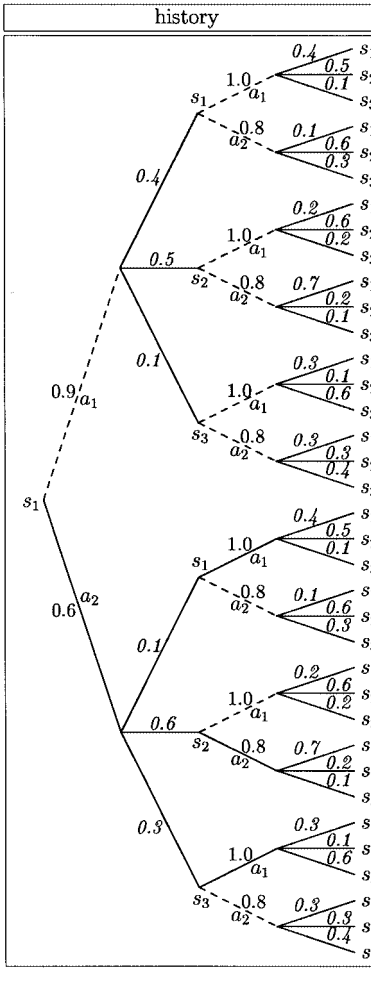
and $2^9 = 512$ second decision functions,

$$\sigma_1 = \begin{pmatrix} \sigma_1(s_1, s_1) & \sigma_1(s_2, s_1) & \sigma_1(s_3, s_1) \\ \sigma_1(s_1, s_2) & \sigma_1(s_2, s_2) & \sigma_1(s_3, s_2) \\ \sigma_1(s_1, s_3) & \sigma_1(s_2, s_3) & \sigma_1(s_3, s_3) \end{pmatrix}.$$

There is a total of $8 \times 512 = 4096$ general policies $\sigma = \{\sigma_0, \sigma_1\}$ for the problem (32).

First, the brute force enumeration in Fig. 1 shows $V^0(s_1) = 0.465$. Similarly, we can calculate the maximum expected values $V^0(s_2)$, $V^0(s_3)$ in Figs. 2 and 3, respectively. (Because of the space limitation we omit Figs. 2 and 3.) Then we have

$$V^0(s_1) = 0.465, \qquad V^0(s_2) = 0.494, \qquad V^0(s_3) = 0.56. \tag{33}$$



history	ter.	path	min	mult.	sub.	total
$s_1 a_1 s_1$	0.5	0.16	0.5	0.08		
$s_1 a_1 s_2$	0.2	0.2	0.2	0.04	0.152	
$s_1 a_1 s_3$	0.8	0.04	0.8	0.032		
$s_1 a_2 s_1$	0.5	0.04	0.5	0.02		
$s_1 a_2 s_2$	0.2	0.24	0.2	0.048	**0.164**	
$s_1 a_2 s_3$	0.8	0.12	0.8	0.096		
$s_2 a_1 s_1$	0.5	0.1	0.5	0.05		
$s_2 a_1 s_2$	0.2	0.3	0.2	0.06	0.19	
$s_2 a_1 s_3$	0.8	0.1	0.8	0.08		0.464
$s_2 a_2 s_1$	0.5	0.35	0.5	0.175		
$s_2 a_2 s_2$	0.2	0.1	0.2	0.02	**0.235**	
$s_2 a_2 s_3$	0.8	0.05	0.8	0.04		
$s_3 a_1 s_1$	0.5	0.03	0.5	0.015		
$s_3 a_1 s_2$	0.2	0.01	0.2	0.002	**0.065**	
$s_3 a_1 s_3$	0.8	0.06	0.8	0.048		
$s_3 a_2 s_1$	0.5	0.03	0.5	0.015		
$s_3 a_2 s_2$	0.2	0.03	0.2	0.006	0.053	
$s_3 a_2 s_3$	0.8	0.04	0.8	0.032		
$s_1 a_1 s_1$	0.5	0.04	0.5	0.02		
$s_1 a_1 s_2$	0.2	0.05	0.2	0.01	**0.036**	
$s_1 a_1 s_3$	0.8	0.01	0.6	0.006		
$s_1 a_2 s_1$	0.5	0.01	0.5	0.005		
$s_1 a_2 s_2$	0.2	0.06	0.2	0.012	0.035	
$s_1 a_2 s_3$	0.8	0.03	0.6	0.018		
$s_2 a_1 s_1$	0.5	0.12	0.5	0.06		
$s_2 a_1 s_2$	0.2	0.36	0.2	0.072	0.204	
$s_2 a_1 s_3$	0.8	0.12	0.6	0.072		**0.465**
$s_2 a_2 s_1$	0.5	0.42	0.5	0.21		
$s_2 a_2 s_2$	0.2	0.12	0.2	0.024	**0.27**	
$s_2 a_2 s_3$	0.8	0.06	0.6	0.036		
$s_3 a_1 s_1$	0.5	0.09	0.5	0.045		
$s_3 a_1 s_2$	0.2	0.03	0.2	0.006	**0.159**	
$s_3 a_1 s_3$	0.8	0.18	0.6	0.108		
$s_3 a_2 s_1$	0.5	0.09	0.5	0.045		
$s_3 a_2 s_2$	0.2	0.09	0.2	0.018	0.135	
$s_3 a_2 s_3$	0.8	0.12	0.6	0.072		

FIG. 1. All two-stage behaviors from s_1 and selection of maximum branch. $V^0(s_1) = \mathrm{Max}_{\sigma_0, \sigma_1} \Sigma_{(x_1, x_2) \in X \times X} \{[r_0(u_0) \wedge r_1(u_1) \wedge r_G(x_2)] p(x_1 \mid s_1, u_0) p(x_2 \mid x_1, u_1)\}$.

The calculation yields, at the same time, the optimal policy $\sigma^* = \{\sigma_0^*(x_0), \sigma_1^*(x_0, x_1)\}$,

$$\sigma_0^*(s_1) = a_2, \qquad \sigma_0^*(s_2) = a_1, \qquad \sigma_0^*(s_3) = a_1$$

$$\sigma_1^*(s_1, s_1) = a_1, \qquad \sigma_1^*(s_2, s_1) = a_2, \qquad \sigma_1^*(s_3, s_1) = a_2$$

$$\sigma_1^*(s_1, s_2) = a_2, \qquad \sigma_1^*(s_2, s_2) = a_2, \qquad \sigma_1^*(s_3, s_2) = a_2$$

$$\sigma_1^*(s_1, s_3) = a_1, \qquad \sigma_1^*(s_2, s_3) = a_1, \qquad \sigma_1^*(s_3, s_3) = a_1.$$

Note that

$$\sigma_1^*(s_1, s_1) \neq \sigma_1^*(s_2, s_1).$$

Thus the optimal policy σ^* is not Markov (but general).

In Fig. 1 we use the following notations:

history $= x_0 \;\; r_0(u_0)/u_0 \;\; p(x_1 \mid x_0, u_0) \;\; x_1 \;\; r_1(u_1)/u_1 \;\; p(x_2 \mid x_1, u_1) \;\; x_2$

ter. = terminal value $= r_G(x_2)$

path = path probability $= p(x_1 \mid x_0, u_0) p(x_2 \mid x_1, u_1)$

min = minimum of the three $= r_0(u_0) \wedge r_1(u_1) \wedge r_G(x_2)$

mult. = path $\times$ min

sub. = subtotal expected value

total = total expected value.

Furthermore, the *italic* face means probability, and the **bold** face denotes a selection of maximum of up expected or down values.

Second, Table I is an arrangement of Figs. 1, 2, and 3, made by selecting all $(8 \times 8 = 64)$ Markov policies $\pi = \{\pi_0, \pi_1\}$. The table lists the corresponding expected value vectors,

$$J^0(\pi) = \begin{pmatrix} J^0(s_1; \pi) \\ J^0(s_2; \pi) \\ J^0(s_3; \pi) \end{pmatrix},$$

where

$$J^0(x_0; \pi) = \sum_{(x_1, x_2) \in X \times X} \{[r_0(u_0) \wedge r_1(u_1) \wedge r_G(x_2)] \times p(x_1 \mid x_0, u_0) p(x_2 \mid x_1, u_1)\}$$

$$u_0 = \pi_0(x_0), \qquad u_1 = \pi_1(x_1), \qquad x_0 = s_1, s_2, s_3$$

$$\pi_0 = \begin{pmatrix} \pi_0(s_1) \\ \pi_0(s_2) \\ \pi_0(s_3) \end{pmatrix} \qquad \pi_1 = \begin{pmatrix} \pi_1(s_1) \\ \pi_1(s_2) \\ \pi_1(s_3) \end{pmatrix}.$$

TABLE I
All Expected Value Vectors $J^0(\pi)$, Where $\pi = \{\pi_0, \pi_1\}$ Is Markov

$\pi_0 \backslash \pi_1$	$\begin{pmatrix} a_1 \\ a_1 \\ a_1 \end{pmatrix}$	$\begin{pmatrix} a_1 \\ a_1 \\ a_2 \end{pmatrix}$	$\begin{pmatrix} a_1 \\ a_2 \\ a_1 \end{pmatrix}$	$\begin{pmatrix} a_1 \\ a_2 \\ a_2 \end{pmatrix}$	$\begin{pmatrix} a_2 \\ a_1 \\ a_1 \end{pmatrix}$	$\begin{pmatrix} a_2 \\ a_1 \\ a_2 \end{pmatrix}$	$\begin{pmatrix} a_2 \\ a_2 \\ a_1 \end{pmatrix}$	$\begin{pmatrix} a_2 \\ a_2 \\ a_2 \end{pmatrix}$
$\begin{pmatrix} a_1 \\ a_1 \\ a_1 \end{pmatrix}$	$\begin{pmatrix} 0.407 \\ 0.434 \\ 0.542 \end{pmatrix}$	$\begin{pmatrix} 0.395 \\ 0.410 \\ 0.470 \end{pmatrix}$	$\begin{pmatrix} 0.452 \\ 0.488 \\ 0.551 \end{pmatrix}$	$\begin{pmatrix} 0.440 \\ 0.464 \\ 0.479 \end{pmatrix}$	$\begin{pmatrix} 0.419 \\ 0.440 \\ 0.551 \end{pmatrix}$	$\begin{pmatrix} 0.407 \\ 0.416 \\ 0.479 \end{pmatrix}$	$\begin{pmatrix} 0.464 \\ 0.494 \\ 0.560 \end{pmatrix}$	$\begin{pmatrix} 0.452 \\ 0.470 \\ 0.488 \end{pmatrix}$
$\begin{pmatrix} a_1 \\ a_1 \\ a_2 \end{pmatrix}$	$\begin{pmatrix} 0.407 \\ 0.434 \\ 0.422 \end{pmatrix}$	$\begin{pmatrix} 0.395 \\ 0.410 \\ 0.390 \end{pmatrix}$	$\begin{pmatrix} 0.452 \\ 0.488 \\ 0.455 \end{pmatrix}$	$\begin{pmatrix} 0.440 \\ 0.464 \\ 0.423 \end{pmatrix}$	$\begin{pmatrix} 0.419 \\ 0.440 \\ 0.419 \end{pmatrix}$	$\begin{pmatrix} 0.407 \\ 0.416 \\ 0.387 \end{pmatrix}$	$\begin{pmatrix} 0.464 \\ 0.494 \\ 0.452 \end{pmatrix}$	$\begin{pmatrix} 0.452 \\ 0.470 \\ 0.420 \end{pmatrix}$
$\begin{pmatrix} a_1 \\ a_2 \\ a_1 \end{pmatrix}$	$\begin{pmatrix} 0.407 \\ 0.373 \\ 0.542 \end{pmatrix}$	$\begin{pmatrix} 0.395 \\ 0.365 \\ 0.470 \end{pmatrix}$	$\begin{pmatrix} 0.452 \\ 0.395 \\ 0.551 \end{pmatrix}$	$\begin{pmatrix} 0.440 \\ 0.387 \\ 0.479 \end{pmatrix}$	$\begin{pmatrix} 0.419 \\ 0.366 \\ 0.551 \end{pmatrix}$	$\begin{pmatrix} 0.407 \\ 0.358 \\ 0.479 \end{pmatrix}$	$\begin{pmatrix} 0.464 \\ 0.388 \\ 0.560 \end{pmatrix}$	$\begin{pmatrix} 0.452 \\ 0.380 \\ 0.488 \end{pmatrix}$
$\begin{pmatrix} a_1 \\ a_2 \\ a_2 \end{pmatrix}$	$\begin{pmatrix} 0.407 \\ 0.373 \\ 0.422 \end{pmatrix}$	$\begin{pmatrix} 0.395 \\ 0.365 \\ 0.390 \end{pmatrix}$	$\begin{pmatrix} 0.452 \\ 0.395 \\ 0.455 \end{pmatrix}$	$\begin{pmatrix} 0.440 \\ 0.387 \\ 0.423 \end{pmatrix}$	$\begin{pmatrix} 0.419 \\ 0.366 \\ 0.419 \end{pmatrix}$	$\begin{pmatrix} 0.407 \\ 0.358 \\ 0.387 \end{pmatrix}$	$\begin{pmatrix} 0.464 \\ 0.388 \\ 0.452 \end{pmatrix}$	$\begin{pmatrix} 0.452 \\ 0.380 \\ 0.420 \end{pmatrix}$
$\begin{pmatrix} a_2 \\ a_1 \\ a_1 \end{pmatrix}$	$\begin{pmatrix} 0.399 \\ 0.434 \\ 0.542 \end{pmatrix}$	$\begin{pmatrix} 0.375 \\ 0.410 \\ 0.470 \end{pmatrix}$	$\begin{pmatrix} 0.465 \\ 0.488 \\ 0.551 \end{pmatrix}$	$\begin{pmatrix} 0.441 \\ 0.464 \\ 0.479 \end{pmatrix}$	$\begin{pmatrix} 0.398 \\ 0.440 \\ 0.551 \end{pmatrix}$	$\begin{pmatrix} 0.374 \\ 0.416 \\ 0.479 \end{pmatrix}$	$\begin{pmatrix} 0.464 \\ 0.494 \\ 0.560 \end{pmatrix}$	$\begin{pmatrix} 0.440 \\ 0.470 \\ 0.488 \end{pmatrix}$
$\begin{pmatrix} a_2 \\ a_1 \\ a_2 \end{pmatrix}$	$\begin{pmatrix} 0.399 \\ 0.434 \\ 0.422 \end{pmatrix}$	$\begin{pmatrix} 0.375 \\ 0.410 \\ 0.390 \end{pmatrix}$	$\begin{pmatrix} 0.465 \\ 0.488 \\ 0.455 \end{pmatrix}$	$\begin{pmatrix} 0.441 \\ 0.464 \\ 0.423 \end{pmatrix}$	$\begin{pmatrix} 0.398 \\ 0.440 \\ 0.419 \end{pmatrix}$	$\begin{pmatrix} 0.374 \\ 0.416 \\ 0.387 \end{pmatrix}$	$\begin{pmatrix} 0.464 \\ 0.494 \\ 0.452 \end{pmatrix}$	$\begin{pmatrix} 0.440 \\ 0.470 \\ 0.420 \end{pmatrix}$
$\begin{pmatrix} a_2 \\ a_2 \\ a_1 \end{pmatrix}$	$\begin{pmatrix} 0.399 \\ 0.373 \\ 0.542 \end{pmatrix}$	$\begin{pmatrix} 0.375 \\ 0.365 \\ 0.470 \end{pmatrix}$	$\begin{pmatrix} 0.465 \\ 0.395 \\ 0.551 \end{pmatrix}$	$\begin{pmatrix} 0.441 \\ 0.387 \\ 0.479 \end{pmatrix}$	$\begin{pmatrix} 0.398 \\ 0.366 \\ 0.551 \end{pmatrix}$	$\begin{pmatrix} 0.374 \\ 0.358 \\ 0.479 \end{pmatrix}$	$\begin{pmatrix} 0.464 \\ 0.388 \\ 0.560 \end{pmatrix}$	$\begin{pmatrix} 0.440 \\ 0.380 \\ 0.488 \end{pmatrix}$
$\begin{pmatrix} a_2 \\ a_2 \\ a_2 \end{pmatrix}$	$\begin{pmatrix} 0.399 \\ 0.373 \\ 0.422 \end{pmatrix}$	$\begin{pmatrix} 0.375 \\ 0.365 \\ 0.390 \end{pmatrix}$	$\begin{pmatrix} 0.465 \\ 0.395 \\ 0.455 \end{pmatrix}$	$\begin{pmatrix} 0.441 \\ 0.387 \\ 0.423 \end{pmatrix}$	$\begin{pmatrix} 0.398 \\ 0.366 \\ 0.419 \end{pmatrix}$	$\begin{pmatrix} 0.374 \\ 0.358 \\ 0.387 \end{pmatrix}$	$\begin{pmatrix} 0.464 \\ 0.388 \\ 0.452 \end{pmatrix}$	$\begin{pmatrix} 0.440 \\ 0.380 \\ 0.420 \end{pmatrix}$

$$J^0(x_0; \pi) = \sum_{(x_1, x_2) \in X \times X} \{[r_0(u_0) \wedge r_1(u_1) \wedge r_G(x_2)] p(x_1 \mid x_0, u_0) p(x_2 \mid x_1, u_1)\}$$

We see that the optimal value vector

$$V^0 = \begin{pmatrix} V^0(s_1) \\ V^0(s_2) \\ V^0(s_3) \end{pmatrix}$$

becomes

$$V^0 = \begin{pmatrix} 0.465 \\ 0.494 \\ 0.56 \end{pmatrix}.$$

Thus, Table I shows that for any Markov policy π,

$$V^0(x_0) > J^0(x_0; \pi) \qquad \text{for some } x_0 \in \{s_1, s_2, s_3\},$$

which completes the proof of Theorem 2.2.

4.2. *Invariant Imbedding Method*

In this subsection we consider the following imbedded problem of (32):

$$\begin{aligned} &\text{Maximize} \quad E\big[\lambda \wedge r_0(u_0) \wedge r_1(u_1) \wedge r_G(x_2)\big] \\ &\text{subject to} \quad \text{(i), (ii)}. \end{aligned} \tag{34}$$

We show that the corresponding recursive equation (14) for the imbedded problem (34) is solved as follows.

First, we have

$$v^2(s_1, \lambda) = \lambda \wedge 0.5, \qquad v^2(s_2, \lambda) = \lambda \wedge 0.2, \qquad v^2(s_3, \lambda) = \lambda \wedge 0.8.$$

Second, the equation for $x_1 = s_1$,

$$v^1(x_1, \lambda) = \max_{u_1 \in \{a_1, a_2\}} \sum_{x_2 \in \{s_1, s_2, s_3\}} v^2\big(x_2, \lambda \wedge r_1(u_1)\big) p(x_2 \,|\, x_1, u_1),$$

yields

$$v^1(s_1, \lambda) = \begin{cases} \lambda & [0, 0.2] \\ 0.5\lambda + 0.1 & [0.2, 0.5] \\ 0.1\lambda + 0.3 & \text{on } [0.5, 0.65] \\ 0.3\lambda + 0.17 & [0.65, 0.8] \\ 0.41 & [0.8, 1] \end{cases}$$

$$\tilde{\pi}_1(s_1, \lambda) = \begin{cases} a_1 \text{ or } a_2 & [0, 0.2] \\ a_1 & \text{on } [0.2, 0.65] \\ a_2 & [0.65, 1]. \end{cases}$$

Similarly, we have

$$v^1(s_2, \lambda) = \begin{cases} \lambda & [0, 0.2] \\ 0.8\lambda + 0.04 & [0.2, 0.5] \\ 0.1\lambda + 0.39 & [0.5, 0.8] \\ 0.47 & [0.8, 1] \end{cases}$$

$$\tilde{\pi}_1(s_2, \lambda) = \begin{cases} a_1 \text{ or } a_2 & [0, 0.2] \\ a_2 & [0.2, 1], \end{cases}$$

$$v^1(s_3, \lambda) = \begin{cases} \lambda & [0, 0.2] \\ 0.9\lambda + 0.02 & [0.2, 0.5] \\ 0.6\lambda + 0.17 & [0.5, 0.8] \\ 0.65 & [0.8, 1] \end{cases}$$

$$\tilde{\pi}_1(s_3, \lambda) = \begin{cases} a_1 \text{ or } a_2 & [0, 0.2] \\ a_1 & [0.2, 1]. \end{cases}$$

Third, the equation

$$v^0(x_0, \lambda) = \max_{u_0 \in \{a_1, a_2\}} \sum_{x_1 \in \{s_1, s_2, s_3\}} v^1(x_1, \lambda \wedge r_0(u_0)) p(x_1 \mid x_0, u_0)$$

yields

$$v^0(s_1, \lambda) = \begin{cases} \lambda & [0, 0.2] \\ 0.8\lambda + 0.04 & [0.2, 0.5] \\ 0.25\lambda + 0.315 & [0.5, 0.6] \\ 0.495 & [0.6, 1] \end{cases}$$

$$\tilde{\pi}_0(s_1, \lambda) = \begin{cases} a_1 \text{ or } a_2 & [0, 0.2] \\ a_2 & [0.2, 1], \end{cases}$$

$$v^0(s_2, \lambda) = \begin{cases} \lambda & [0, 0.2] \\ 0.76\lambda + 0.048 & [0.2, 0.5] \\ 0.2\lambda + 0.328 & [0.5, 0.65] \\ 0.24\lambda + 0.302 & [0.65, 0.8] \\ 0.494 & [0, 8.1] \end{cases}$$

$$\tilde{\pi}_0(s_2, \lambda) = \begin{cases} a_1 \text{ or } a_2 & [0, 0.2] \\ a_1 & [0.2, 1], \end{cases}$$

and

$$v^0(s_3, \lambda) = \begin{cases} \lambda & [0, 0.2] \\ 0.77\lambda + 0.046 & [0.2, 0.5] \\ 0.4\lambda + 0.231 & \text{on } [0.5, 0.65] \\ 0.46\lambda + 0.192 & [0.65, 0.8] \\ 0.56 & [0.8, 1] \end{cases}$$

$$\tilde{\pi}_0(s_3, \lambda) = \begin{cases} a_1 \text{ or } a_2 & \text{on } [0, 0.2] \\ a_1 & [0.2, 1]. \end{cases}$$

Therefore, the optimization problem (34) has the maximum expected values

$$v^0(s_1, 1) = 0.465, \qquad v^0(s_2, 1) = 0.494, \qquad v^0(s_3, 1) = 0.56.$$

Thus, we have obtained the optimal expected value $v^0(x_0, 1)$, which is coincident with $V^0(x_0)$ in (33), obtained by the brute force enumeration method in Section 4.1:

$$v^0(x_0, 1) = V^0(x_0) \qquad \text{for } x_0 \in \{s_1, s_2, s_3\}.$$

Now let us construct through (15) an optimal general policy $\tilde{\sigma} = \{\tilde{\sigma}_0, \tilde{\sigma}_1\}$ from the Markov policy $\tilde{\pi} = \{\tilde{\pi}_0, \tilde{\pi}_1\}$ for the imbedded process (34).

First, we have the first decision function:

$$\tilde{\sigma}_0(s_1) = \tilde{\pi}_0(s_1, 1) = a_2, \qquad \tilde{\sigma}_0(s_2) = \tilde{\pi}_0(s_2, 1) = a_1,$$
$$\tilde{\sigma}_0(s_3) = \tilde{\pi}_0(s_3, 1) = a_1.$$

Second, we see that the second decision function in (15),

$$\sigma_1(x_0, x_1) = \pi_1(x_1, \lambda_1) = \pi_1(x_1, \lambda_0 \wedge r_0(x_0, u_0)), \qquad u_0 = \pi_0(x_0, \lambda_0),$$

reduces in our data to

$$\tilde{\sigma}_1(s_1, x_1) = \tilde{\pi}_1(x_1, r_0(a_2)) = \tilde{\pi}_1(x_1, 0.6)$$
$$\tilde{\sigma}_1(s_2, x_1) = \tilde{\pi}_1(x_1, r_0(a_1)) = \tilde{\pi}_1(x_1, 0.9)$$
$$\tilde{\sigma}_1(s_3, x_1) = \tilde{\pi}_1(x_1, r_0(a_1)) = \tilde{\pi}_1(x_1, 0.9).$$

This yields

$$\tilde{\sigma}_1(s_1, s_1) = a_1, \qquad \tilde{\sigma}_1(s_2, s_1) = a_2, \qquad \tilde{\sigma}_1(s_3, s_1) = a_2$$
$$\tilde{\sigma}_1(s_1, s_2) = a_2, \qquad \tilde{\sigma}_1(s_2, s_2) = a_2, \qquad \tilde{\sigma}_1(s_3, s_2) = a_2$$
$$\tilde{\sigma}_1(s_1, s_3) = a_1, \qquad \tilde{\sigma}_1(s_2, s_3) = a_1, \qquad \tilde{\sigma}_1(s_3, s_3) = a_1.$$

Thus, we have through invariant imbedding obtained an optimal policy $\tilde{\sigma}$, which is not Markov but general. Of course, the optimal policy $\tilde{\sigma}$ is coincident with σ^* obtained by the enumeration method in Section 4.1.

REFERENCES

1. R. E. Bellman, "Dynamic Programming," Princeton Univ. Press, Princeton, NJ, 1957.
2. R. E. Bellman and E. D. Denman, "Invariant Imbedding," Lecture Notes in Operation Research and Mathematical Systems, Vol. 52, Springer-Verlag, Berlin, 1971.
3. R. E. Bellman and L. A. Zadeh, Decision-making in a fuzzy environment, *Management Sci*. **17** (1970), B141–B164.
4. D. P. Bertsekas, "Dynamic Programming and Stochastic Control," Academic Press, New York, 1976.
5. D. P. Bertsekas and S. E. Shreve, "Stochastic Optimal Control," Academic Press, New York, 1978.
6. D. Blackwell, Discounted dynamic programming, *Ann. Math. Statist*. **36** (1965), 226–235.
7. E. V. Denardo, Contraction mappings in the theory underlying dynamic programming, *SIAM Rev*. **9** (1968), 165–177.
8. E. V. Denardo, "Dynamic Programming: Models and Applications," Prentice-Hall, Englewood Cliffs, NJ, 1982.
9. A. O. Esogbue and R. E. Bellman, Fuzzy dynamic programming and its extensions, *TIMS/Studies Management Sci*. **20** (1984), 147–167.
10. R. Hartley, L. C. Thomas, and D. J. White (eds.), Recent development in Markov decision processes, *in* "Proceedings of an International Conference on Markov Decision Processes," Academic Press, New York, 1980.
11. K. Hinderer, "Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter," Lecture Notes in Operation Research and Mathematical Systems, Vol. 33, Springer-Verlag, Berlin, 1970.
12. R. A. Howard, "Dynamic Programming and Markov Processes," MIT Press, Cambridge, MA, 1960.
13. S. Iwamoto, From dynamic programming to bynamic programming, *J. Math. Anal. Appl.* **177** (1993), 56–74.
14. S. Iwamoto, On bidecision processes, *J. Math. Anal. Appl*. **187** (1994), 676–699.
15. S. Iwamoto, Associative dynamic programs, *J. Math. Anal. Appl*. **201** (1996), 195–211.
16. S. Iwamoto and T. Fujita, Stochastic decision-making in a fuzzy environment, *J. Oper. Res. Soc. Japan* **38** (1995), 467–482.
17. J. Kacprzyk, Decision-making in a fuzzy environment with fuzzy termination time, *Fuzzy Sets and Systems* **1** (1978), 169–179.
18. D. M. Kreps, Decision problems with expected utility criteria, I, *Math. Oper. Res*. **2** (1977), 45–53.
19. D. M. Kreps, Decision problems with expected utility criteria, II, stationarity, *Math. Oper. Res*. **2** (1977), 266–274.
20. E. S. Lee, "Quasilinearization and Invariant Imbedding," Academic Press, New York, 1968.
21. L. G. Mitten, Composition principles for synthesis of optimal multi-stage processes, *Oper. Res*. **12** (1964), 610–619.
22. G. L. Nemhauser, "Introduction to Dynamic Programming," Wiley, New York, 1966.
23. E. Porteus, An informal look at the principle of optimality, *Management Sci*. **21** (1975), 1346–1348.

24. E. Porteus, Conditions for characterizing the structure of optimal strategies in infinite-horizon dynamic programs, *J. Optim. Theory Appl.* **36** (1982), 419–432.
25. M. L. Puterman (ed.), Dynamic programming and its applications, *in* "Proceedings of the International Conference on Dynamic Programming and Its Applications," Academic Press, New York, 1978.
26. M. L. Puterman, "Markov Decision Processes: Discrete Stochastic Dynamic Programming," Wiley, New York, 1994.
27. M. Sniedovich, "Dynamic Programming," Dekker, New York, 1992.
28. N. L. Stokey and R. E. Lucas, Jr., "Recursive Methods in Economic Dynamics," Harvard Univ. Press, Cambridge, MA, 1989.
29. D. J. White, "Dynamic Programming," Holden-Day, San Francisco, 1969.
30. D. J. White, "Finite Dynamic Programming," Wiley, New York, 1978.
31. P. Whittle, "Optimization over Time," Vols. I and II, "Dynamic Programming and Stochastic Control," Wiley, New York, 1982, 1983.